AN IMAGE PROCESSING APPROACH TO COMPUTING DISTANCES BETWEEN RNA SECONDARY STRUCTURES DOT PLOTS

By

Tor Ivry

Shahar Michal

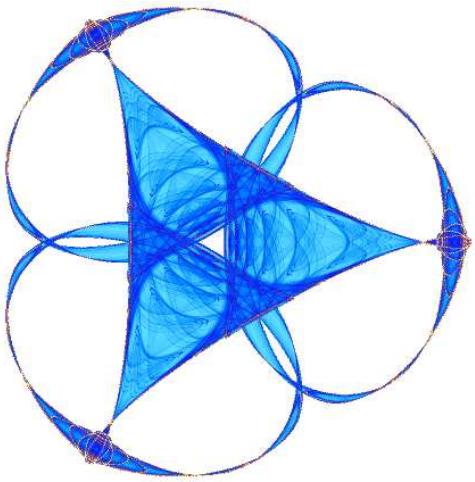
Danny Barash

and

Guillermo Sapiro

IMA Preprint Series # 2163

(March 2007)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

400 Lind Hall

207 Church Street S.E.

Minneapolis, Minnesota 55455–0436

Phone: 612-624-6066 Fax: 612-626-7370

URL: http://www.ima.umn.edu

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
An Image Processing Approach to Computing Distances Between RNA Secondary Structures Dot Plots (PREPRINT)	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Minnesota,Institute for Mathematics and its Applications,207 Church Street SE,Minneapolis,MN,55455-0436**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Background: Computing the distance between two RNA secondary structures can contribute in understanding the functional relationship between them. When used repeatedly, such a procedure may lead to finding a query RNA structure of interest in a database of structures. Several methods are available for computing distances between RNAs represented as strings or graphs, but none utilize the RNA representation with dot plots. Since dot plots are essentially digital images, there is a clear motivation to devise an algorithm for computing the distance between dot plots based on image processing methods. Results: We have developed a new metric dubbed 'DoPloCompare', which compares two RNA structures. The method is based on comparing dot plot diagrams that represent the secondary structures. When analyzing two diagrams and motivated by image processing, the distance is based on a combination of histogram correlations and a geometrical distance measure. We illustrate the procedure by an application that utilizes this metric on RNA sequences in order to locate peculiar point mutations that induce significant structural alternations relative to the wild type predicted secondary structure. The method was tested on several RNA sequences with known secondary structures to a±rm their prediction, as well as on a data set of ribosomal pieces. These pieces were computationally cut from a ribosome for which an experimentally derived secondary structure is available, and on each piece the prediction conveys similarity to the experimental result. The new algorithm shows benefit when compared to standard methods used for assessing the distance similarity between two RNA secondary structures. Conclusions: Inspired by image processing, we have managed to provide a conceptually new and potentially beneficial metric for comparing two RNA secondary structures, and illustrated it on an application that utilized the measurement to detect conformational rearranging point mutations on an RNA sequence.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**28**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

An Image Processing Approach to Computing Distances Between RNA Secondary Structures Dot Plots

Tor Ivry[1] , Shahar Michal[1] , Danny Barash[1] and Guillermo Sapiro[2*]

[1] Department of Computer Science, Ben-Gurion University, Israel.
[2] Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, USA.

Email: Tor Ivry - ivryt@cs.bgu.ac.il; Shahar Michal - mshaha@cs.bgu.ac.il; Danny Barash - dbarash@cs.bgu.ac.il; Guillermo Sapiro - guille@umn.edu;

*Corresponding author

Abstract

Background: Computing the distance between two RNA secondary structures can contribute in understanding the functional relationship between them. When used repeatedly, such a procedure may lead to finding a query RNA structure of interest in a database of structures. Several methods are available for computing distances between RNAs represented as strings or graphs, but none utilize the RNA representation with dot plots. Since dot plots are essentially digital images, there is a clear motivation to devise an algorithm for computing the distance between dot plots based on image processing methods.

Results: We have developed a new metric dubbed 'DoPloCompare', which compares two RNA structures. The method is based on comparing dot plot diagrams that represent the secondary structures. When analyzing two diagrams and motivated by image processing, the distance is based on a combination of histogram correlations and a geometrical distance measure. We illustrate the procedure by an application that utilizes this metric on RNA sequences in order to locate peculiar point mutations that induce significant structural alternations relative to the wild type predicted secondary structure. The method was tested on several RNA sequences with known secondary structures to affirm their prediction, as well as on a data set of ribosomal pieces. These pieces were computationally cut from a ribosome for which an experimentally derived secondary structure is available, and on each piece the prediction conveys similarity to the experimental result. The new algorithm shows benefit when compared to standard methods used for assessing the distance similarity between two RNA secondary structures.

Conclusions: Inspired by image processing, we have managed to provide a conceptually new and potentially beneficial metric for comparing two RNA secondary structures, and illustrated it on an application that utilized the measurement to detect conformational rearranging point mutations on an RNA sequence.

Background

In the past several years, interesting novel RNAs were discovered that carry a diverse array of functionalities. By now, it is well known that RNAs are considerably involved in mediating the synthesis of proteins, regulating cellular activities, and exhibiting enzyme-like catalysis and post-transcriptional activities. In many of these cases, knowledge of the RNA secondary structure can be helpful to understanding its functionality.

The importance of the secondary structure of RNAs presents a need for tools that rely on comparing two RNA secondary structures, which may indicate a functional commonality or divergence between them. These tools can usually accompany secondary structure prediction packages by energy minimization such as Mfold [1] and the Vienna package [2]. Calculating the distance between RNA structures have been approached by several methods, some of which are based on the edit distance of a tree representation of the RNA secondary structure elements [3–5]. An edit distance on homeomorphically irreducible trees (HITs) [6] was one of the original proposals for a comparison method. A different method was based on the alignment of a string representation of the secondary structures [7, 8], where parenthesis represent the base-pairs, and another symbol represents unpaired nucleotides [5]. This representation is known as the dot-bracket representation. All aforementioned comparison methods were implemented as part of the Vienna RNA package [2, 5]. More recent suggestions for RNA secondary structure comparisons include the use of context free grammars [9], and a more general edit distance under various score schemes [10, 11]. A method for a rapid similarity analysis using the Lempel-Ziv algorithm was suggested in [12]. Another method uses the second eigenvalue of the tree graph representation for the structures comparison, [13], and was later integrated into the RNAMute, [14], Java tool, which we will use for our application illustration.

Certain RNA molecules can act as conformational switches, by alternating between two states, and thereby changing their functionality [15–19]. RNA conformational switching was found to be involved in cell processes such as mRNA transcription, translation, splicing, synthesis and regulation. Given a

thermodynamically stable RNA structure, we can try to predict a conformational rearranging point mutation by traversing all possible single point mutations of a sequence and locate the most significant ones, in terms of secondary structure difference [20]. RNAMute [14] and RDMAS [21] are tools that attempt to perform such predictions and are based on energy minimization methods [1, 2]. The RNAMute mutation analysis tool, [14], includes RNAdistance from [2, 5]: the RNA edit distance of the dot bracket representation as a fine-grain comparison method, and the edit distance of the Shapiro representation, [3, 4], as a coarse-grain comparison method.

Here, we propose an alternative distance measure, motivated by image processing and pattern recognition. The new metric is based on an analysis of the dot plot diagrams of the secondary structures, and uses histogram based correlation and plane group distance to calculate the similarity between the diagrams. The measure combines both fine and coarse elements in the structure and can offer an alternative method to the aforementioned distance measures, with a critical advantage in applications that use energy and probability dot plots for the analysis of secondary structures. We have developed a stand-alone procedure called DoPloCompare, which receives two RNA structures as an input, and calculates their similarity grade using our new distance measure algorithm. In order to illustrate our metric, we have built an application that uses the DoPloCompare procedure to predict the most significant point-mutation in a given sequence that will alter its secondary structure to form a new conformation. Our system uses a user defined external folding program. In the results of this paper it relies on the folding predictions of Mfold, [1], and the Vienna RNA package [5], both using the expanded energy rules by [22] to predict the folding of RNA sequences.

In the following sections we will describe the new procedure DoPloCompare, its application details, and the results obtained when applying the system on three well-studied structures [23–25]. These systems were already examined in [13] in this context. Additionally, we apply DoPloCompare on a ribosomal small RNA sequences data set extracted from [26], and discuss its contribution alongside commonly used routines such as the RNAdistance [5].

DoPloCompare - Comparing Two RNA Secondary Structures

The basis for our algorithm is the fact that a base–pairing indicator dot plot diagram is a sound representation of the RNA secondary structure, as will be detailed in the next Section. In general, a dot plot is a matrix comparison of two sequences (or one with itself) and is prepared by sliding a window of user-defined size along both sequences. If the two sequences within that window match with a precision set

by the mismatch limit, a dot is placed in the middle of the window signifying a match [27]. In the case of RNA sequences, we assume that a similarity between dot plot diagrams of two sequences is a good criterion for similarity between the secondary structures of those sequences.

Given two dot plot diagrams of two secondary structures, we would like to develop a distance grade that best indicates how well the secondary structures attached to the diagrams resemble each other. When two structures are similar, we require that the distance between their representing dot plot diagrams to be small, and alternatively, when the structures are different, we require that the distance will increase.

Observations

Two main observations served as motivation in establishing the distance calculation formula. The first is that similar secondary structures will maintain matching dot plot diagrams with dots in the same or in close positions. Obviously, two secondary structures will look alike if all or most of the base-pairing couples will be located in the same or in proximal places in the sequences. The second observation is that two secondary structures will count as similar if both the number and order of the elements they contain are the same [13]. For example, two RNA structures with four stems can be considerably different if the first structure is arranged as a one elongated structure containing a bulge and three loops (see Figure 1B), while the second includes a bulge, a multi-branch loop, and two additional set-loops that branch out of the multi-branch loop (see Figure 1A). From the second observation, we concluded that the calculation should also reflect the overall arrangement of elements in the secondary structure, and the groups of points in the dot plot diagrams accordingly.

Distance Calculation

Taking into account the two observations, we have developed the following distance grade formula.

Let O be the dot plot diagram of the original sequence representing its secondary structure.

Let M be the dot plot diagram of the mutated sequence representing its secondary structure.

Then:

$$Distance\_Grade(O, M) = \frac{Dist(O, M)}{Corr(O, M)} \tag{1}$$

Where $Corr$ stands for Correlation and $Dist$ stands for Distance. For the Correlation part we used the histograms method as detailed in the Methods Section. In our implementation, we used a 4–dimensional histograms correlation:

$$Corr(O,M) =$$

$$\sqrt{Xc(O,M) \times Yc(O,M) \times Dc(O,M) \times Ic(O,M)} \quad (2)$$

Where:

- Xc(O,M) is the correlation grade (see Equation 4 in Methods) between the vectors that sums all the points on each X column of the matrix
- Yc(O,M) is the correlation grade between the vectors that sums all the points on each Y row of the matrix
- Dc(O,M) is the correlation grade between the vectors that sums all the points on each Diagonal SW-NE
- Ic(O,M) is the correlation grade between the vectors that sums all the points on each Inverse Diagonal SE-NW

For the distance part we used the RMS distance as explained in the Methods Section.

Formulas Explanation

The histogram correlation compares the locations of every p_i and p_j under the best matching shift, where p_i is a pixel in the original sequence's dot plot diagram, and p_j is a pixel in the mutated sequence's dot plot diagram. However, in some cases small differences in the locations of the pixels between the original and the mutated dot plot diagrams, reduces the correlation grade. Literally, the grade is reduced for every pixel in the original dot plot that is not placed on the same exact location as a pixel in the mutated dot plot. For this reason, we introduce a distance measure between the dot plot diagrams, in addition to the histogram correlation.

The distance measure is more tolerant to small differences and represent overall proximity between the sets of points. Moreover, if a pixel in the original dot plot is not placed on top of a pixel in the compared dot plot, the correlation grade will be reduced equally, regardless of the distance between the pixels, while the distance measure will be reduced in a direct proportion to the distance between the pixels.

DoPloCompare Program Flow

DoPloCompare receives two RNA secondary structures as input, either in a dot bracket notations or as two ct files (produced by Mfold [1]). The main flow of the algorithm is made of three parts:

1. Build the dot plot matrix from the secondary structures.
2. Compare the two structures using formula (1) for the distance grade. In order to normalize the distance grade, it is divided by the length of the sequences.
3. Output the distance grade.

Building the Dot Plot Matrix

Taking the simple matrix characteristics (described in the Methods Section), one can easily build such a matrix by traversing a folding option received as an output of any folding program, and for every base-pairing nucleotides couple in the sequence set the matching matrix cell value to 1 (other cell values will be set to 0).

Application for Finding the Most Significant Point Mutation

The system is based on both histograms and geometry as the core comparing mechanism between the original sequence secondary structure and all the possible point mutations' folding variants. The algorithm is composed of two major parts: pre-processing and main comparing mechanism. The pseudo–code of the algorithm is given here:

```
Most_Significant_Mutation ( Original_Sequence )
BEGIN
    Original_Matrix:= Built matrix
        from the folding of Original_Sequence;
    Max_Grade:=0;
    Max_Sequence:=Original_Sequence;
    WHILE ( Mutated_Sequence := Next
        point mutation of Original_Sequence )
    BEGIN
        Mutated_Matrix:=Built matrix from the
            folding of Mutated_Sequence;
        Grade:=Distance grade between
            Original_Matrix and Mutated_Matrix;
        If ( Grade > Max_Grade )
        BEGIN
            Max_Grade:=Grade;
            Max_Sequence:=Mutated_Sequence;
        END
    END
    Return Max_Sequence;
END.
```

System Parameters

The system has several parameters, including:

- Folding program – either MFOLD or Vienna's RNAsubopt.
- Number of suboptimal folding options to be considered by the algorithm.
- Geometric distance measure to be used – either RMS or Hausdorff [28] distances. The default measure is RMSD.

Pre-processing

The pre-processing part is divided to three steps (each is described in detail in the Methods Section):

1. Create all single–point–mutations in the original sequence.
2. Fold the mutated sequences using the folding program of choice.
3. From the folding program's output, we build a dot plot like matrix.

Main Comparing Mechanism

The mutated and original secondary structures' representing dot plot matrices are being compared using the DoPloCompare application (see 'DoPloCompare' section). Each mutated sequence's dot plot matrix receives a distance grade, which represents its similarity to the original sequence's representing matrix.

Output

At this stage, the algorithm finds the dot plot with the highest distance grade, i.e., the dot plot with the greatest difference from the dot plot diagram of the original sequence. This dot plot represents the secondary structure of one of the suboptimal folding options of a mutated sequence. The algorithm reports this sequence, along with additional data:

1. A representation of the secondary structure - either a dot-bracket in the case of RNAsubopt or a ct file in the case of Mfold.
2. The location of the point mutation and the replaced nucleotide (e.g., G15U).
3. The dot-plot-like matrix of the mutated sequence.

In addition, for user convenience, the secondary structure and the dot-plot-like matrix elements of the original sequence are also attached.

Results

In order to test our system capabilities, we applied it to three test cases that were used in [13] and compared our results to the aforementioned work. Additionally, we tested our system on a data set of ribosomal RNA pieces.

Wild Type Sequences

We will describe the results for three well-studied RNA sequences that were used in [13] for a bioinformatics proof of concept. It is worthwhile noting that we are looking for the mutation with the largest structural difference from the wild type, while in [13] the ultimate goal was to look for a mutation that can lead to a bistable conformation. We successfully locate mutations that lead to a folding rearrangement with large difference from the wild type structure, and that are similar to the ones found in [13]. In addition to the second eigenvalue classification, we specifically compare our results to RNAdistance's dot bracket edit distance grade, which was mentioned but not directly used for comparison in [13]. RNAdistance was later integrated into RNAMute [14].

Leptomonas collosoma

The first sequence is the spliced leader RNA from *Leptomonas collosoma* which was studied by LeCuyer and Crothers [23], where they experimentally demonstrated a mutation induced RNA switch. In this test case, our system reported a structure with one double strand segment and a hairpin. This structure is of larger difference from the optimal wild type folding than the one reported in [13] that contains a bulge and a hairpin. We assume that this difference emerges from the different folding parameters, because the second eigenvalue of our result is also 1.0. A supporting fact for the latter is that when taking the largest RNAdistance grade, we obtain the same mutation and suboptimal folding as ours. The results are presented in Figure 2.

P5abc subdomain

The second sequence is the P5abc subdomain of the *tetrahymena thermophila* ribozyme that was studied by Wu and Tinoco [24]. The results for the second sequence are found in Figure 1. In this test case, our

system predicted the mutation G15C, which was also reported in [13] as a solution. When testing the P5abc subdomain with Mfold, both G15C and G15U produced the same dot plot matrix in one of their suboptimal folding options, thus receiving the same similarity grade. The mutation C22G produced a very similar matrix, with a somewhat lower similarity grade. In this case, the largest RNAdistance grade was received in the mutated structure of A4C, which is more similar to the original structure than our results. Both the A4C mutation and the original structure contain a multi-branch loop, while our reported mutation's structure does not.

Hepatitis delta virus

The third sequence is taken from human hepatitis delta virus ribozyme that was studied by Lazinski et al. [25], for its regulation of self-cleavage activity. The results for the third sequence are found in Figure 3. In this test case, our system predicted the C31G mutation. The structure induced by this mutation is similar to the one in [13]. The U40G that was suggested in their research [25] maintained a similarity grade that was very close to the grade of our system result. In [25], the authors mention the existence of eight possible mutations that provide the desired non-linear effect in the ribozyme structure, and this may explain the variation. The largest RNAdistance score was recorded in a highly similar structure to the one found by our system.

Ribosomal Data-set

We have generated a data set of small RNA sequences, containing fragments that were cut from the rRNA of the *thermus thermophilus* [26]. This data set was built in order to test our system and compare its results to the RNAdistance results. Labels for the data set can be found in the Supplementary Information file. Out of the 21 RNA sequences in the data set, 16 produced the same exact mutation and structure as the ones received by comparing the edit distance of the dot bracket representation of the folded structures. Two sequences produced different mutations but highly similar structures to the results from RNAdistance. Regarding the remaining three sequences, there was a difference between our system result and the largest RNAdistance result:

1. Our proposed structure for the $E\_(89)$ is different than the structure with the largest RNAdistance, but it is non-obvious to determine which one of them is more significant, both of the mutations alter the structure with respect to the original structure, as observed in Figure 4(A).

2. Our proposed structure for the $E\_(86, 87)$ is quite similar to the structure with the largest RNAdistance. However, both the RNAdistance structure and the original structure contains an extra loop. Thus, it can be argued that our proposed structure is less similar to the original one, as observed in Figure 4(B).

3. Our proposed structure for the B_(1052-1107) is less similar to the original structure than the structure with the largest RNAdistance. Both the original and RNAdistance's structures contain a branch that is not present in our system's result, as can be observed in Figure 4(C).

The ribosomal data set results are summarized in Table 1. Labelings for the sequences that are used in Table 1 are reported in the Supplementary Information file.

Discussion and Future Work

We have described a method to compare two RNA secondary structures, and to assign a grade to this comparison based on the similarity of their representing dot matrices. We have adopted this method to predict the most significant point mutation for a given sequence in terms of its structural effect on the wildtype, and provided good results in comparison to other known methods.

We have compared our application results to the commonly used RNAdistance module provided in the Vienna package [2, 5], and the classification by the second eigenvalue that was provided for three example test cases in [13]; the first result, from *Leptomonas collosoma*, was less similar to the original structure than the one predicted in [13] (i.e., in this test case our system surpassed). However, we assume this difference is partly caused by the different folding program and parameters. For the second result, the P5abc subdomain, our system predicted a mutation that was proposed in [13], and on the final result, from the hepatitis delta virusoid, we have predicted a very similar structure to the one found by the second eigenvalue method. Overall our system matched or even outranked the second eigenvalue method results.

Concerning the results for the ribosomal data set, which were compared to RNAdistance's results: the results were identical in 16 out of the 21 RNA sequences, 2 sequences produced different mutations but highly similar structures to the results from RNAdistance, and for the remaining 3 sequences, there was a difference between our system results and the largest RNAdistance results. However, for these three sequences, we argue that our results presented mutated structures with less similarity to the original structures, when comparing to the structures with the largest RNAdistance. Thus, overall our system outperformed RNAdistance results in at least some of the cases.

The distance measure presented in this article, DoPloCompare, has several advantages with respect to previously suggested techniques (most commonly used are the ones described in [5]):

- The measure is used with the dot plot representation, whereas to the best of our knowledge no other measure was suggested beforehand for this type of representation. Probability and energy dot plots have an increased potential to be used even more in the future, in cases where a more sophisticated analysis is needed besides inspecting the predicted secondary structures. The measure is inversely proportional to the similarity (or proportional to the dissimilarity) between the structures being compared.
- The metric combines coarse and fine-grain characteristics, provided by the distance measure and the correlation respectively, and thus balances both the distance between the nucleotides and the structural elements (e.g., hairpin, loop, etc.) in the compared structures.
- DoPloCompare is easily tuned with regard to the distance function (Hausdorff, RMS, etc.), the correlation algorithm (histograms correlations, traditional correlation, etc.) and their combination.
- DoPloCompare can receive the structures as input from a list of popular folding programs' output files, such as Mfold and the Vienna RNA package.
- DoPloCompare is incorporated into an application that predicts the most conformational rearranging point mutations, and provides good results in comparison to known methods.

There are a number of avenues we propose to pursue in the future for the extension of DoPloCompare and the presented application:

- DoPloCompare: operation on more sophisticated dot plots that contain more information (e.g., probability and / or energy values). Our technique using histogram correlation and RMS distance permits for potential extensions that will utilize numerical values contained within dots, much like in the case of digital images.
- DoPloCompare: integrate into the RNAMute mutation analysis tool [14].
- Finding the most conformational rearranging mutations: extend to handle deletions, insertions, and multiple–point mutations using efficiency considerations.

Conclusions

We have provided a new beneficial technique to compare secondary structures of RNA sequences. The technique is robust and can be used as a baseline for other RNA structure based applications.

Methods

RNA Suboptimal Solutions

In order to make predictions based on an RNA secondary structure, we used the RNAsubopt [29] available in the Vienna RNA package, a program that predicts all suboptimal secondary structures of a given sequence based on thermodynamics and base-pairing rules [22]. Alternatively, we can use the suboptimal solutions calculated by Mfold. RNAsubopt, like many other RNA folding approaches, uses a free energy minimization procedure. It is expected that the native fold of the sequence is close to the minimum free energy (mfe) structure. We are interested in all suboptimal solutions because in nature RNA folds into a suboptimal structure (and also because of limitations of thermodynamic models), which may cause the mfe structure to be different than the native fold. For a given sequence, RNAsubopt calculates all suboptimal secondary structures within an energy range above the minimum free energy. It outputs the suboptimal structures—sorted by mfe—in a dot-bracket notation, followed by the energy in kcals/mol. Originally, a different method for calculating suboptimal solutions was devised by Zuker [30], and is used in Mfold.

Creating the Point Mutations

In order to create all the possible single point mutations for a given sequence, we simply traverse along the sequence and for each position i do:

Let $N_1,\ N_2$ *and* N_3 be the three possible nucleotides which are different than the nucleotide in position i.

Let *SEQ(j,k)* denote the subsequence starting in position j in the original sequence and ending at position k (in case $k < j$ return an empty sequence).

Return:

$$SEQ(1, i-1) \circ N_1 \circ SEQ(i+1, m) \bigcup$$
$$SEQ(1, i-1) \circ N_2 \circ SEQ(i+1, m) \bigcup$$
$$SEQ(1, i-1) \circ N_3 \circ SEQ(i+1, m)$$

Where m is the original sequence length.

Dot Plot Diagrams

A dot plot is a diagram comprised of dots on two axes. Each of the axis represents some sort of data. A dot in location (x, y) represents some measure between the location x in the X-data axis and location y in the Y-data axis. For example, the axis can represent two sentences, and the dots can represent the locations where the sentence on the X-axis and the sentences on the Y-axis contain the same word.
In biology, dot plots are often utilized for representing alignments between sequences. Specifically in RNA, a dot plot is often used as an image representation of an optimal base-pairing between any two nucleotides in the RNA sequence, based on minimum free energy consideration. Both Mfold [1] and the Vienna RNA package [5] present dot plots as part of their standard outputs, but instead of dots they use squares. Mfold presents dot plot diagrams based on the minimum free energy of the suboptimal folding options of the sequence, where each folding option squares are painted with a different color. Vienna-RNA, on the other hand, presents a different dot plot diagram where each square in the diagram represents the probability of a base-pairing in that location in the sequence; the larger the probability, the larger the representing square. In our approach, we compare each folding option separately, and require a separate dot plot diagram for each suboptimal solution (as opposed to Mfold's dot plot, for example). To comply with this constraint, we created a simplified dot-plot-like matrix with the following properties:

1. Let LEN be the length of the sequence being observed, then the matrix is of two dimensions, and of size $LEN \times LEN$.

2. The matrix cell (i,j) can contain either one of the values {0,1} where 1 means that i match j in the current folding option and 0 otherwise.

Giving the fact that if i matches j, j will also match i, clearly the matrix is symmetric along the diagonal.

Histograms

Histograms have been widely and very successfully used in image processing and shape analysis. Although originally they were used to study the data statistics, they have recently been found to be critical for identification, recognition, and distance computations as well, e.g., [31, 32]. Such histograms constitute the building block of most state of-the-art shape identification and classification systems. Moreover, it has been recently shown that under very general conditions, histograms can uniquely identify a shape with extremely high probability [33]. This provides a very clear motivation to consider histograms for RNA secondary structure analysis, as suggested in this paper.

In order to explain the "Dist" and "Corr" components of Equation (1) in more detail, we will first concentrate on "Corr" (which is, in our case, the Cross_Correlate expressed in Equation 4). Next, in the subsection about the distance between groups of points in the plane, we will concentrate on "Dist" (which is, in our case, the RMSD expressed in Equation 5).

In this manuscript we are using normalized cross-correlation between two one-dimensional vectors.

Cross correlation is a standard method of estimating the degree to which two vectors are correlated.

Consider two vectors, $X(i)$ and $Y(i)$, where $i = 0, 1, 2...N-1$.

The cross correlation $Corr$ at delay d is defined as:

$$Corr(d) = \frac{\sum_i [(X(i) - MX) \times (Y(i-d) - MY)]}{\sqrt{\sum (X(i) - MX)^2 \times \sum (Y(i-d) - MY)^2}} \tag{3}$$

Where MX and MY are the means of the corresponding series, and $d = 0, 1, 2, ...N-1$ represents all the possible delays.

In this paper we refer to the cross correlation between X and Y as:

$$Cross\_Correlate(X, Y) = Max_d(Corr(d)) \tag{4}$$

Where $Corr(d)$ is as defined in Equation 3.

In order to build a one–dimensional series vector from the two–dimensional matrix that represents the original Dot Plot diagram, we traverse the diagram, each time on a specific axis, and sum all the values on that axis (e.g. sum all the columns on the X axis, or sum all the rows on the Y axis). In this manner we obtain a one–dimensional vector for each axis, which can be correlated to the matching axis vector of the second matrix that represents the mutated Dot Plot diagram (see example in Figure 5).

The Cross-Correlation grade will be maximal when the two compared vectors are identical, or contain identical areas. We have used this feature in our assumptions, as explained in the DoPloCompare Section under the distance calculation subsection.

Distance Between Groups of Points in the Plane

The matching and analysis of geometric features is an important problem that arises in various computational areas, e.g., computer vision and pattern matching . In general, we are given two sets of points A and B, and we wish to determine how much they resemble each other (for more information see [34]). Usually we can apply certain transformation on one of the sets, e.g., translate, scale and/or rotate, in order to be matched with the other set as closely as possible.

In order to measure affinity, various measure functions have been devised. Two such common measures are the Hausdorff distance [34] and the root mean square distance (RMS) [35–37]. Note that the Hausdorff distance has also been popular in image processing [28].

In this paper we use the RMS measure (e.g., Dist-RMS in Equation 5), but the system can be easily adapted to use the Hausdorff measure or any other measure. No alignment between the groups is performed, after several trials have shown no difference if an alignment is added, and therefore the alignment procedure was removed for performance considerations.

The Root Mean Square distance for a set B from set A is:

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{a\in A}\|a - N_B(a)\|^2} \tag{5}$$

Where n is the size of group A and $N_B(a)$ is the nearest neighbor of point a in group B.

The mark $\|$ in this context refers to the Euclidean norm.

The measure simply sums and normalizes the distances between each point in A to its nearest neighbor in set B. Clearly, when the two sets lie on top each other, the RMS score will be 0, Alternatively, for sets of different spreading in the plane the RMS distance will increase.

RMS distance between groups of points uses nearest neighbor queries in order to find the point from the other group from which to calculate each point's distance. In order to calculate nearest neighbor queries we implemented a version of planar Voronoi diagram [38], with pre-process time of $O(n)$, which answers nearest neighbor queries in $O(\log n)$ for a group of n locations in the plane. We chose not to further discuss Voronoi diagram as its implementation and use has no influence on the system output but only on the algorithm run-time.

In our approach, we look for the distance between groups of dots in the base-pairing plane, i.e., we look for the RMS distance between two dot plot diagrams which is explained in detail in the "DoPloCompare" Section under the distance calculation subsection.

Base-pairing Distance

As a baseline method for comparing two secondary structures we used RNAdistance, which is also part of the Vienna-RNA package. It reads RNA secondary structures and calculates a "base-pair distance" given by the number of base pairs present in one structure—but not the other.

We use this method as a measure of success in identifying the largest distance between the original sequence and the mutated sequence.

We compare our results to RNAdistance fine-grain method where two structures in dot-bracket notations are being compared.

Authors Contributions

TI and SM worked on the software design, carried our development and implementation, and participated in drafting the manuscript. DB and GS conceived the study, coordinated the software design and drafted the manuscript.

Acknowledgements

Special thanks are reserved to Boaz Rosenberg who provided some insightful ideas on implementing the correlation algorithm. We acknowledge the support of the Lynn and William Frankel Center for Computer Sciences. GS is supported by DARPA, NSF, and ONR.

References

1. Zuker M: **Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction.** *Nucleic Acids Res.* 2003, **31**:3406–3415.
2. Hofacker I: **Vienna RNA Secondary Structure Server.** *Nucleic Acids Res.* 2003, **31**:3429–3431.
3. Shapiro BA: **An Algorithm for Comparing Multiple RNA Secondary Structures.** *CABIOS* 1988, **4**:381–393.
4. Shapiro BA, Zhang K: **Comparing Multiple RNA Secondary Structures Using Tree Comparisons.** *CABIOS* 1993, **33**:309–318.
5. Hofacker I, Fontana W, Stadler P, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures.** *Monatsh. Chem.* 1994, **125**:167–188.
6. Fontana W, Konings D, Stadler P, Schuster P: **Statistics of RNA Secondary Structures.** *Biopolymers* 1993, **33**:1389–1404.
7. Hogeweg P, Hesper B: **Energy Directed Folding of RNA Sequences.** *Nucleic Acids Res.* 1984, **12**:67–74.
8. Konings DAM, Hogeweg P: **Pattern Analysis of RNA Secondary Structure Similarity and Consensus of Minimal-Energy Folding.** *J. Mol. Biol.* 1989, **207**:597–614.
9. Holmes I, Rubin GM: **Pairwise RNA Structure Comparison with Stochastic Context-Free Grammars.** In *Proceedings of the Pac. Symp. Biocomputing*, 7, World Scientific 2002:163–174.
10. Jiang T, Lin G, Ma B, Zhang K: **A General Edit Distance between RNA Structures.** *J. Comput. Biol.* 2002, **9**:371–388.
11. Hochsmann M, Voss B, Giegerich R: **Pure multiple RNA secondary structure alignments: a progressive profile approach.** *IEEE/ACM Trans Comput Biol Bioinform.* 2004, **1**:53–62.
12. Liu N, Wang T: **A Method for Rapid Similarity Analysis of RNA Secondary Structures.** *BMC Bioinformatics* 2004, **7**(493).
13. Barash D: **Second Eigenvalue of the Laplacian Matrix for Predicting RNA Conformational Switch by Mutation.** *Bioinformatics* 2004, **20**:1861–1869.
14. Churkin A, Barash D: **RNAmute: RNA Secondary Structure Mutation Analysis Tool.** *BMC Bioinformatics* 2006, **7**(201).

15. Schultz EA, Bartel D: **One Sequence, Two Ribozymes.** *Science* 2000, **289**:448–452.

16. Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA, Nudler E: **Sensing Small Molecules by Nascent RNA: a Mechanism to Control Transcription in Bacteria.** *Cell* 2002, **111**:747–756.

17. Winkler W, Nahvi A, Breaker RR: **Thiamine Derivatives Bind Messenger RNAs Directly to Regulate Bacterial Gene Expression.** *Nature* 2002, **419**:952–956.

18. Biebricher CK, Diekmann S, Luce R: **Structural Analysis of Self-Replicating RNA Synthesis by Q_β Replicase.** *J. Mol. Biol.* 1982, **154**:629–648.

19. Biebricher CK, Luce R: **In Vitro Recombination and Terminal Elongation of RNA by Q_β Replicase.** *EMBO J.* 1992, **11**:5129–5135.

20. Barash D: **Deleterious Mutation Prediction in the Secondary Structure of RNAs.** *Nucleic Acids Res.* 2003, **31**:6578–6584.

21. Shu W, Bo X, Liu R, Zhao D, Zheng Z, Wang S: **RDMAS: a web server forRNA deleterious mutation analysis.** *BMC Bioinformatics* 2006, **7**(404).

22. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure.** *J. Mol. Biol.* 1999, **288**:911–940.

23. LeCuyer KA, Crothers DM: **Kinetics of an RNA Molecular Switch.** In *Proceedings of the National Academy of Sciences, USA*, 91 1994:3373–3377.

24. Wu M, Tinoco I: **RNA Folding Causes Secondary Structure Rearrangement.** In *Proceedings of the National Academy of Sciences, USA*, 95 1998:11555–11560.

25. Lazinski DW, Taylor JM: **Regulation of Hepatitis Delta Virus Ribozymes: to Cleave or not to Cleave?** *RNA* 1995, **1**:225–233.

26. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF: **Crystal structure of the ribosome at 5.5 A resolution.** *Science* 2001, **292**:883–896.

27. Maizel J, Lenk R: **Enhanced graphic matrix analysis of nucleic acid and protein sequences.** *Proc. Natl. Acad. Sci. USA* 1981, **78**:7665–7669.

28. Huttenlocher D, Klanderman G, Rucklidge W: **Comparing Images Using the Hausdorff Distance?** *IEEE Trans. Patt. Anal. Mach. Intell.*. 1993, **15**(9):850–863.

29. Wuchty S, Fontana W, Hofacker I, Schuster P: **Complete Suboptimal Folding of RNA and the Stability of Secondary Structures.** *Biopolymers* 1999, **49**:145–165.

30. Zuker M: **On Finding All Suboptimal Foldings of an RNA Molecule.** *Science* 1989, **244**:48–52.

31. Funkhouser T, Kazhdan M, Min P, Shilane P: **Shape-Based Retrieval and Analysis of 3D Models?** *Comm. of the ACM* 2005, **48**(6):58–64.

32. Lowe DG: **Distinctive Image Features from Scale-Invariant Keypoints?** *Int. J. Comput. Vision* 2004, **60**(2):91–110.

33. Boutin M, Kemper G: **Which Point Configurations are Determined by the Distribution of their Pairwise Distances?** *Int. J. Compt. Geometry and Appl.* 2007, **In Press**.

34. Alt H, Guibas L: **Discrete Geometric Shapes: Matching, Interpolation, and Approximation.** In *Handbook of Computational Geometry*, 1st edition. Edited by Sack JR, Urrutia J, Amsterdam: Elsevier 1999:121–153.

35. Besl PJ, McKay ND: **A Method for Registration of 3-d Shapes.** *IEEE Trans. Pattern Analysis and Machine Intelligence* 1992, **14**:239–256.

36. Gelfand N, Ikemoto L, Rusinkiewicz S, Levoy M: **Geometrically Stable Sampling for the ICP Algorithm.** In *Proc. International Conference on 3D Digital Imaging and Modeling, Canada* 2003:260–267.

37. Har-Peled S, Sadri B: **How Fast is the K-Means Method?** *Algorithmica* 2005, **41**:185–202.

38. Huttenlocher D, Kedem K, Sharir M: **The Upper Envelope of Voronoi Surfaces and its Applications.** In *Proceedings of the Ann. Symp. on Computational Geometry*, 7 1991:194–203.

Figures

Figure 1 - P5abc Subdomain

The predicted most significant mutation for the P5abc subdomain in the group I intron ribozyme of the *T. thermophila*. (A) Wild-type folded structure along with its representing dot plot matrix. The computed RNAfold global minimum energy is dG = -26.6. (B) The mutated folded structure with the largest distance grade from DoPloCompare (DP) = 0.102. The RNAdistance grade for this structure (Rdist) = 28. The computed RNAfold global minimum energy is dG = -18.8. (C) The mutated folded structure with the largest RNAdistance grade (Rdist) = 32. The DoPloCompare grade (DP) = 0.070. The computed RNAfold global minimum energy is dG = -22.2 kcals/mole.

Figure 2 - L. Collosoma

The predicted most significant mutation for the spliced leader RNA from *L.collosoma*. (A) Wild-type folded structure along with its representing dot plot matrix. The computed RNAfold global minimum energy is dG = -10.7. (B) The mutated folded structure with the largest distance grade from DoPloCompare (DP) = 0.102. The largest RNAdistance grade was also recorded for this structure (Rdist) = 52. The computed RNAfold global minimum energy is dG = -8.1 kcals/mole.

Figure 3 - Delta Virusoid

The predicted most significant mutation for the virusoid sequence from Hepatitis delta virus. (A) Wild-type folded structure along with its representing dot plot matrix. The computed RNAfold global minimum energy is dG = -68.6. (B) The mutated folded structure with the largest distance grade from DoPloCompare (DP) = 0.023. The RNAdistance grade for this structure (Rdist) = 60. The computed RNAfold global minimum energy is dG = -67.5. (C) The mutated folded structure with the largest RNAdistance grade (Rdist) = 62. The DoPloCompare grade (DP) = 0.022. The computed RNAfold global minimum energy is dG = -63.7 kcals/mole.

Figure 4 - Ribosomal Data-set Differences

Three examples from the ribosomal data set that produced differences between our system proposed structure and the structure with the largest RNAdistance. (A) The original structure of item $E\_(89)$ from the ribosomal data set (left) along with our system resulted structure (center) and the structure with the largest RNAdistance (right). (B) The same results set for $E\_(86, 87)$. (C) The results set for $B\_(1052 - 1107)$.

Figure 5 - Sum Vectors for Dot-Plot Matrix

A 10×10 dot plot diagram sample, along with its four representing sum vectors:

- The 'X Sum Vector' which sums all the dots values along the X axis of the diagram.
- The 'Y Sum Vector' which sums all the dots values along the Y axis of the diagram.
- The 'Diagonal SW-NE Sum Vector' which sums all the dots along the SW-NE diagonal of the diagram.
- The 'Inverse Diagonal SE-NW Sum Vector' which sums all the dots along the SE-NW inverse diagonal of the diagram.

Where 'Position' refers to a position along the scanned axis, and 'Magnitude' stands for the summed pixel values at that position. The four vectors are being compared to other dot plot diagram's vectors in the process of correlation.

Tables

Table 1 - Ribosomal Data-Set

This table summarizes the results for the ribosomal data set, comparing our system results to the results with the largest RNAdistances. In the fourth column we present our system's predicted mutation. When the resulted mutations are identical to RNAdistance, they are presented in bold face. (A) Marks the 2 sequences with a different mutation but similar structure. (B) Marks the 3 sequences with different secondary structure (Refer also to Figure 4).

Table 1: Ribosomal Data-Set

Index in the data set	Sequence name	Length (nt.)	Our predicted mutation	Mutation with largest RNAdistance [5]
1	A_(765-816)	52	**G7C**	G7C
2	E_(68)	46	**C28G**	C28G
3	A_(1241-1296)	56	G33C$^{(A)}$	G32C
4	A_(820-879)	53	**C4A**	C4A
5	A_(588-651)	64	**G38C**	G38C
6	A_(995-1045)	55	**G41C**	G41C
7	B_(1052-1107)	56	G55A$^{(B)}$	C28U
8	B_(589-668)	82	**G37U**	G37U
9	A_(136-227)	93	**G10U**	G10U
10	A_(1113-1187)	74	**G60U**	G60U
11	B_(865-911)	46	**C38G**	C38G
12	E_(2676-2731)	57	**C3A**	C3A
13	E_(99,100,101)	79	**G9C**	G9C
14	E_(90,91,92)	76	G44A$^{(A)}$	G43A
15	E_(89)	43	G36C$^{(B)}$	A23C
16	D_(8,9,10)	53	**C36G**	G31U
17	A_(1420-1480)	56	**G47C**	G47C
18	A_(240-286)	47	**U5C**	U5C
19	A_(442-492)	41	**G24U**	G24U
20	E_(65,66)	57	**U22A**	U22A
21	E_(86,87)	39	G29A$^{(B)}$	G5C

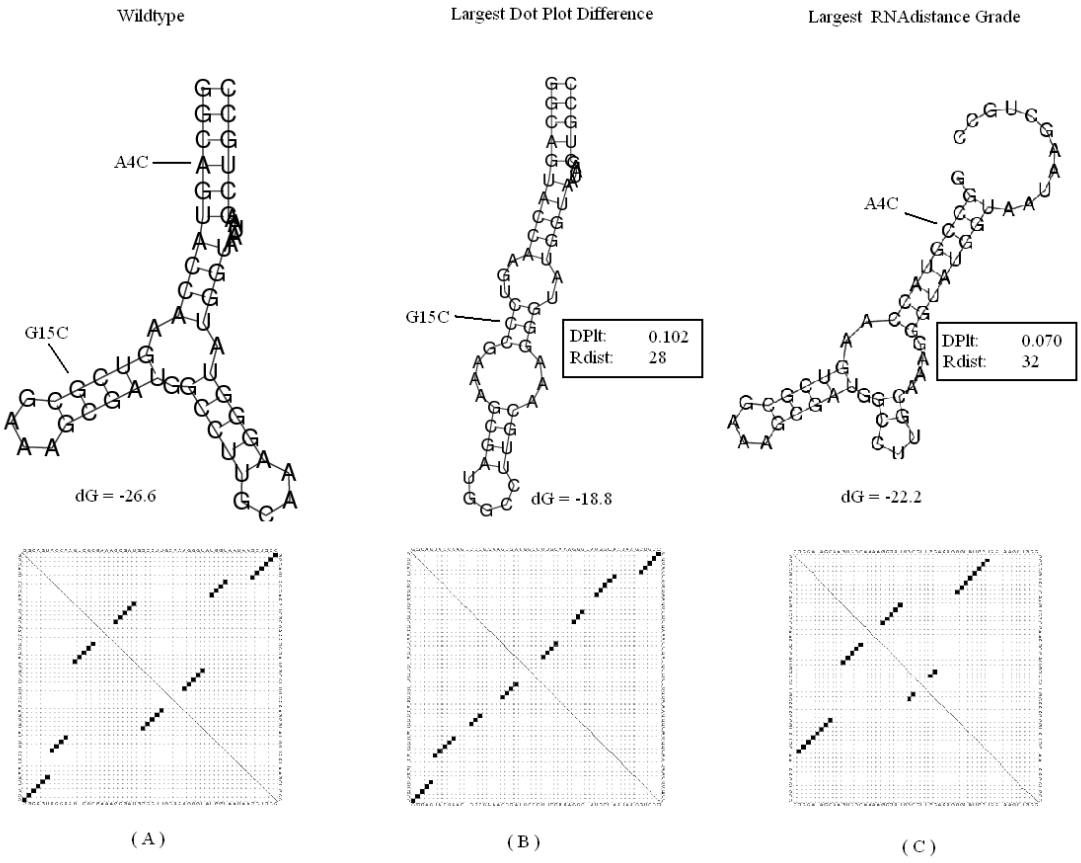


Figure 1: Testcase involving the P5abc subdomain of the *tetrahymena thermophila* ribozyme

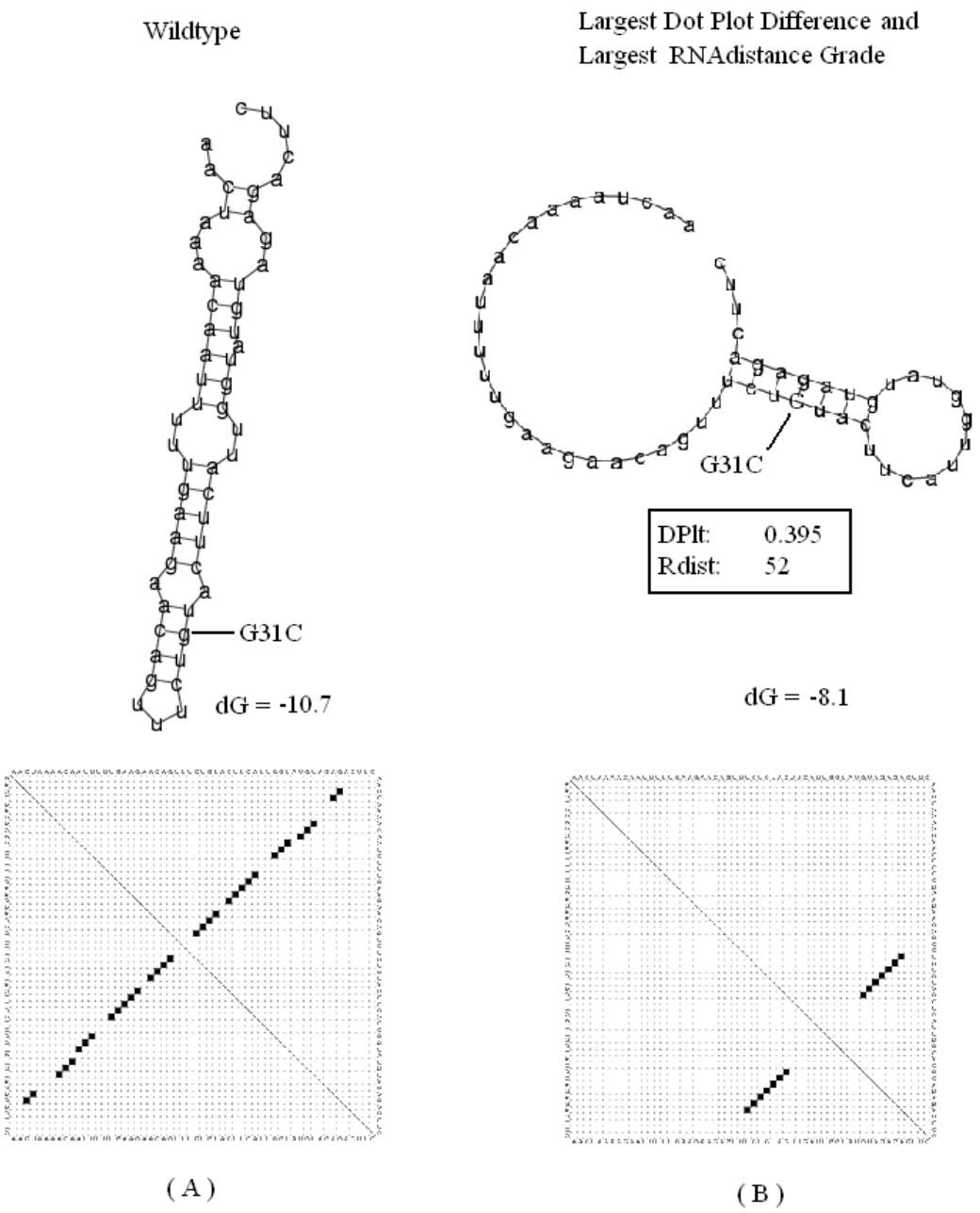


Figure 2: Testcase involving the *L. Collosoma* spliced leader RNA

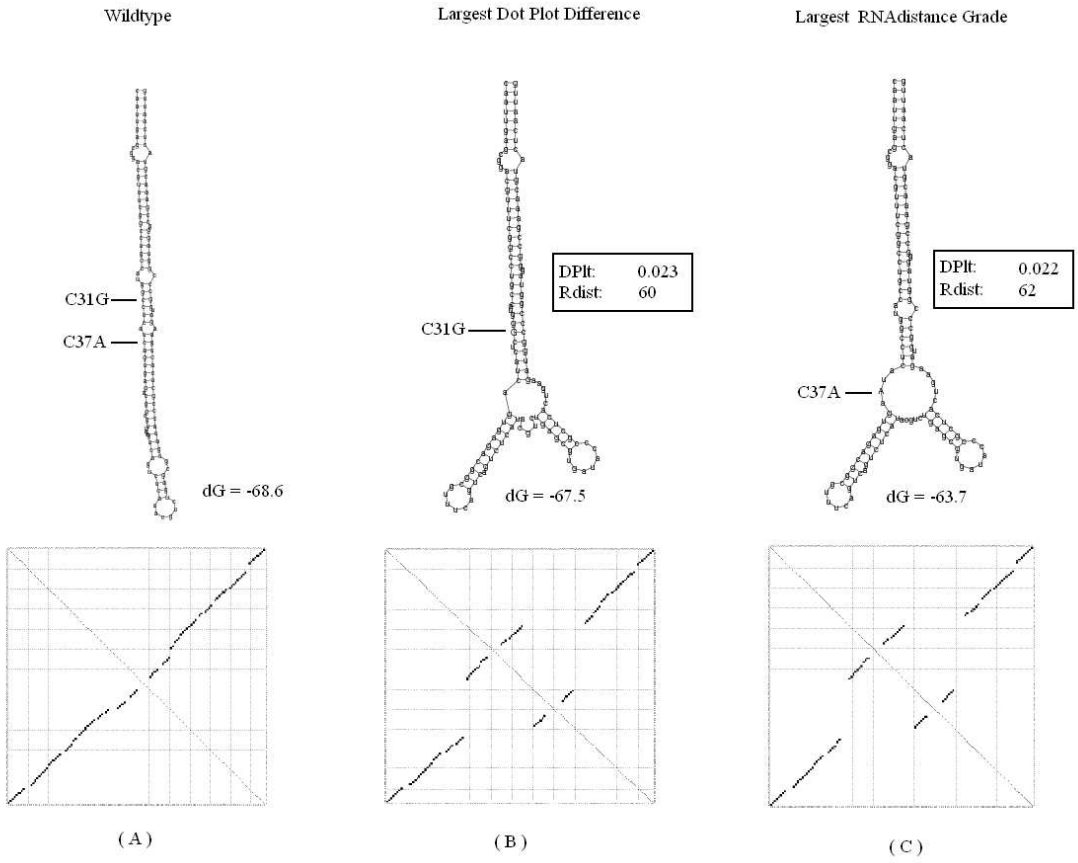


Figure 3: Testcase involving the hepatitis delta virusoid

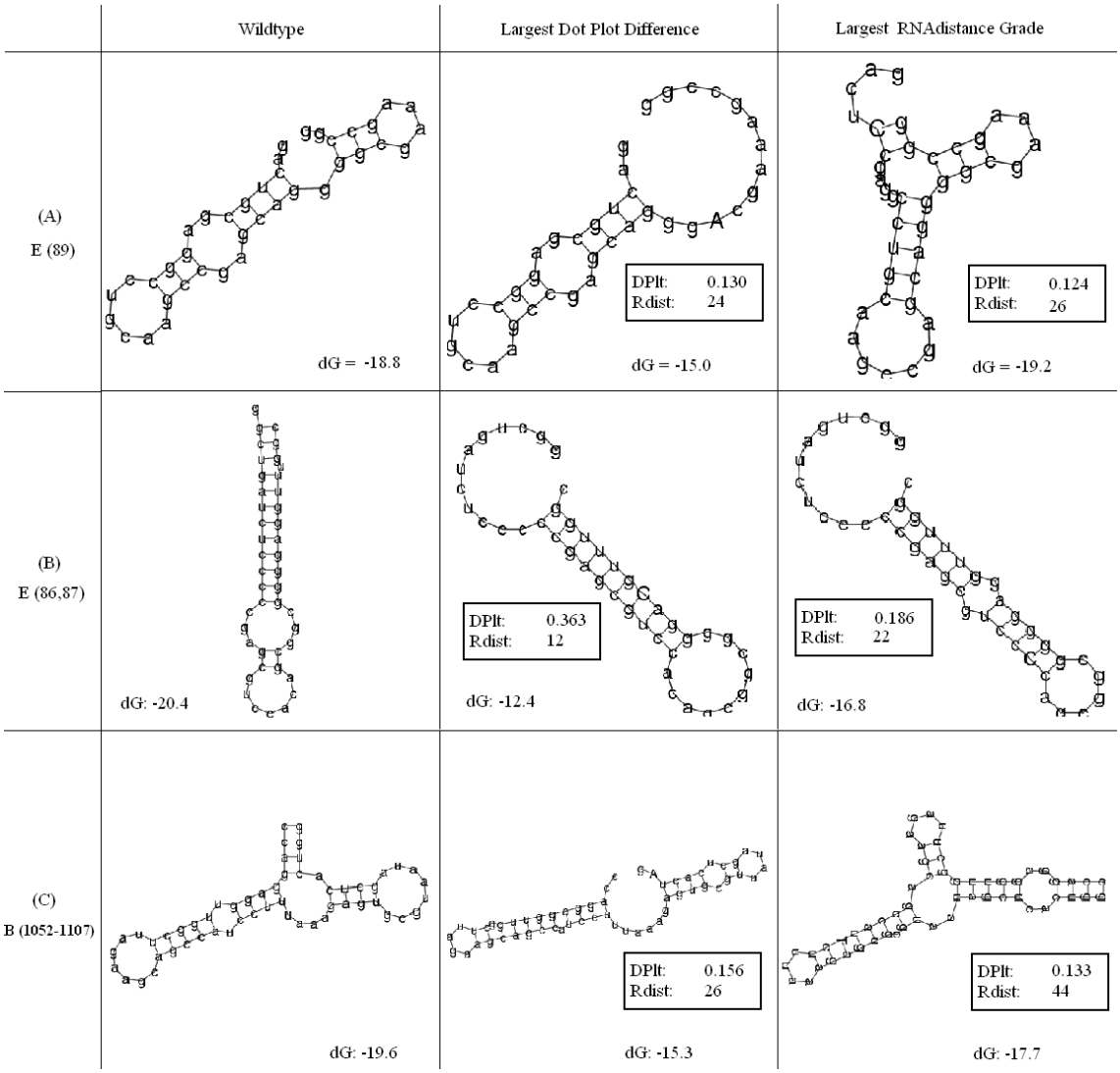


Figure 4: Ribosomal Data-set Differences

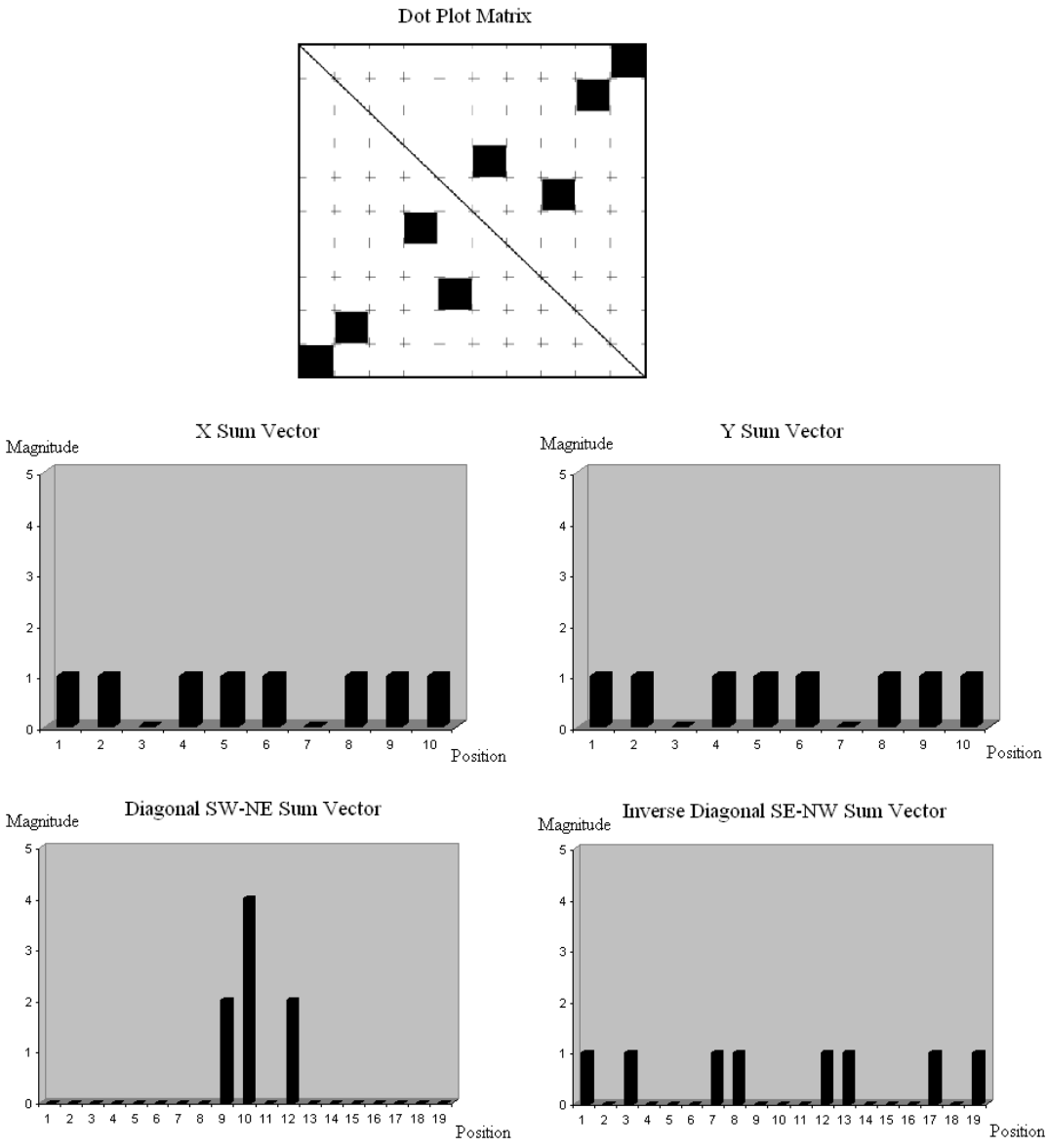


Figure 5: Sum Vectors for Dot-Plot Matrix

An Image Processing Approach to Computing Distances Between RNA Secondary Structures Plots

Supplementary Data

The following Dataset was used in the Results section of the article:

- Dataset of the Ribosomal RNA fragments of *Thermus thermophilus HB8* based on [1] containing the following 21 fragments:

>Entry:A_(765-816) Length:52 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
gaaagcgugggagcaaaccggauuagauacccggguaguccacgcccuaaa

>Entry:E_(68) Length:46 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
ccggaaggucaaggggaggggugcaagccccgaaccgaagccccgg

>Entry:A_(1241-1296) Length:56 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
gcccacuacaaagcgaugccacccggcaacggggagcuaaucgcaaaaaggugggc

>Entry:A_(820-879) Length:53 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
gcgcgcuaggucucugggucuccuggggccgaagcuaacgcguuaagcgcgc

>Entry:A_(588-651) Length:64 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
gccuggggcgucccaugugaaagaccacggcucaaccgugggggagcgugggauacgcucaggc

>Entry:A_(995-1045) Length:55 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
augcuagggaacccgggugaaagccuggggugccccgcgaggggagcccuagcac

>Entry:B_(1052-1107) Length:56 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
ccaggagguuggcuuagaagcagccauccuuuaaagagugcguaauagcucacugg

>Entry:B_(589-668) Length:82 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
cacggucgugggcgagcuuaagccguugaggcggaggcguagggaaaccgaguccgaacagggcgucuaguccgcggccgug

>Entry:A_(136-227) Length:93 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
ccggaagagggggacaacccggggaaacucgggcuaaucccccauguggacccgccccuuggggugugucccaaagggcuuug
cccgcuuccgg

>Entry:A_(1113-1187) Length:74 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
ccccgccguuaguugccagcgguucggccgggcacucuaacgggacugcccgcgaaagcgggaggaaggagggg

>Entry:B_(865-911) Length:46 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
cacugauagggcuagggggcccaccagccuaccaaacccugucaaa

>Entry:E_(2676-2731) Length:57 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
cgcaccucugguuucccagcugucccuccaggggcagaagcuggguagccaugugcg

>Entry:E_(99,100,101) Length:79 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
ggacccgggaagaccacccgguggaugggccggggguguaagcgccgcgaggcguugagccgaccggucccaaucgucc

>Entry:E_(90,91,92) Length:76 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
cggcucgucgcauccuggggcugaagaaggucccaaggguuggggcuguucgcccauuaaagcggcacgcgagcugg

>Entry:E_(89) Length:43 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
ggcugaucucccccgagcguccacagcggcggggagguuuggc

>Entry:D_(8,9,10) Length:53 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
aaugggggaacccggccggcgggaacgccggucaccgcgcuuuugcgcggggg

>Entry:A_(1420-1480) Length:56 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
cgggcucuacccgaagucgccgggagccuacgggcaggcgccgaggguagggcccg

>Entry:A_(240-286) Length:47 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
cccaucagcuaguuggugggguaauggcccaccaaggcgacgacggg

>Entry:A_(442-492) Length:41 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
cccgggacgaaacccccgacgaggggacugacgguaccggg

>Entry:E_(65,66) Length:57 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
acuguuuaccaaaaacacagcucucugcgaacucguaagaggagguauagggagcga

>Entry:E_(86,87) Length:39 Origin:rRNA of the *Thermus thermophilus* [AC:NC_006461]
gacugcgaggccugcaagccgagcaggggcgaaagccgg

1. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF: Crystal structure of the ribosome at 5.5 A resolution. Science 2001, 292:883–896.